회귀분석팀

6팀

심은주
진수정
문병철
이수정
임주은

INDEX

0

지난 주 복습

- **회귀분석이란?**

설명변수 X와 종속변수 Y의 관계를 표현한 식을 찾는 것

단순선형회귀

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

설명변수를 p개로 확장

다중선형회귀

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} + \varepsilon_i$$

- 유의성 검정

F-test: 모델 전체에 대한 검정

H_0: $\beta_1 = \beta_2 = \cdots = \beta_p = 0$

H_1: $\beta_1, \beta_2, \ldots, \beta_p$ 중 적어도 하나는 0 이 아니다.

if 기각되지 않는다면?

- $y = \beta_0 + \varepsilon$ ($\because \beta_1 = \cdots = \beta_p = 0$)
- 회귀식이 아무런 의미가 없음을 의미!

t-test: 개별 회귀계수의 유의성을 검정

H_0: $\beta_j = 0$ 다른 변수들이 적합된 상태에서 x_j는 통계적으로 유의하지 않다

H_1: $\beta_j \neq 0$ 다른 변수들이 적합된 상태에서 x_j는 통계적으로 유의하다

if $|t_j| \geq t_{n-p-1;\alpha/2}$

- 귀무가설 기각
- 다른 변수들이 적합된 상태에서, x_j는 통계적으로 유의한 변수

- **적합성 검정**

결정 계수: 총 변동에서 회귀식이 설명하는 부분

결정계수

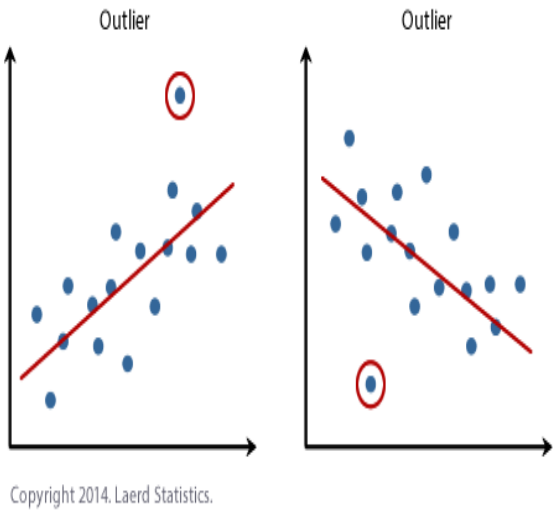
$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

변수의 개수가 다른 두 회귀모형의
직접적인 비교가 어렵다는 점을 해결

수정결정계수

$$R_a^2 = \frac{SSR/p}{SST/(n-1)} = 1 - \frac{SSE/(n-p-1)}{SST/(n-1)}$$

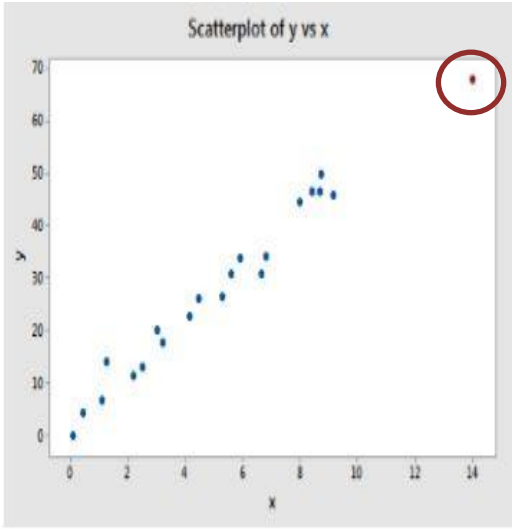
- 데이터 진단



이상치

표준화 잔차가

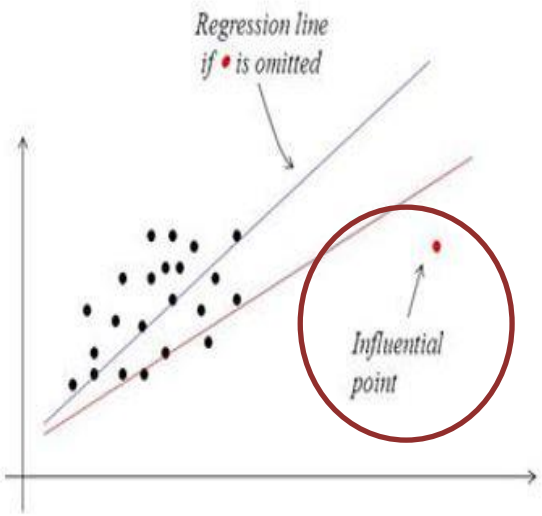
$|r_i| > 3$인 값



지렛값

표준화했을 때,

x 기준에서 절대값이 큰 값!



영향점

회귀직선의 기울기에

상당한 영향을 주는 점

- **로버스트(Robust) 회귀란?**

  건장한, 탄탄한

  이상치의 영향력을 크게 받지 않는 회귀모형

- **로버스트(Robust) 회귀 종류**

Median Regression	Huber's M-estimation	Least Trimmed Square

1

회귀 가정

- 회귀분석의 가정

$$y_i = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p \ , \ \varepsilon \sim NID(0, \sigma^2)$$

1. 식 자체가 X 변수들의 '선형결합'으로 이루어짐
2. 오차는 정규분포(N)를 따름
3. 오차들은 서로 독립적(ID)
4. 오차의 평균은 0, 분산은 σ^2

- **모델의 선형성**

반응변수 y와 예측변수 $x_1 \sim x_p$의 관계가 선형

$$y_i = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

선형 변환 O

$$y = \beta_0 e^{\beta_1 x_1}$$

$$y^* = \log y$$
$$= log\beta + \beta_1 log x_1 = \beta *_0 + \beta_1 x_1$$

변환된 x 를 새로운 x로 취급한다면

선형결합

선형 변환 X

$$y = \frac{\beta_1 x}{\beta_0 + x}$$

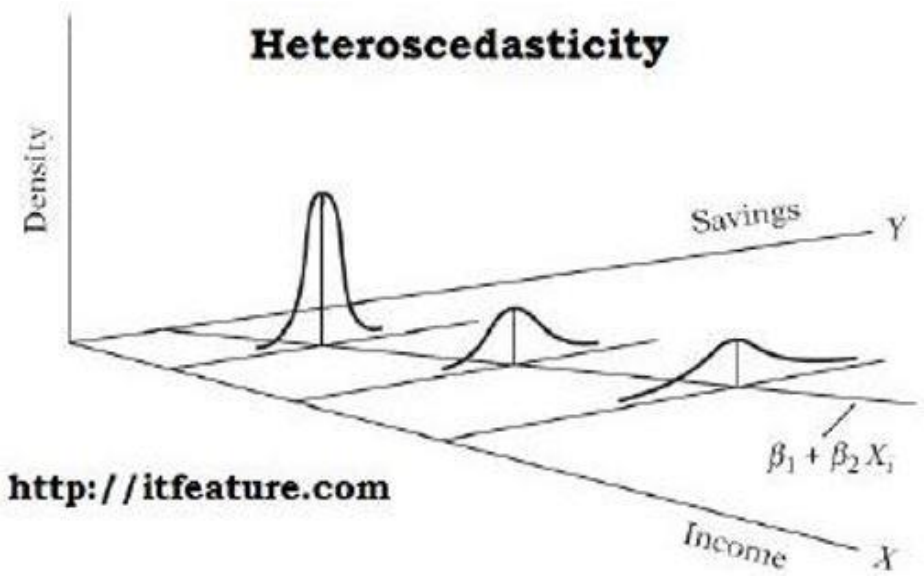
x 와 y 를 변환해도

선형 결합 꼴로 만들 수 없는 모델은

비선형 모델

- **오차의 등분산성**

오차의 분산이 σ^2로 동일한 것



if 가정이 깨진다면?

전체적인 회귀계수의 분산 커짐

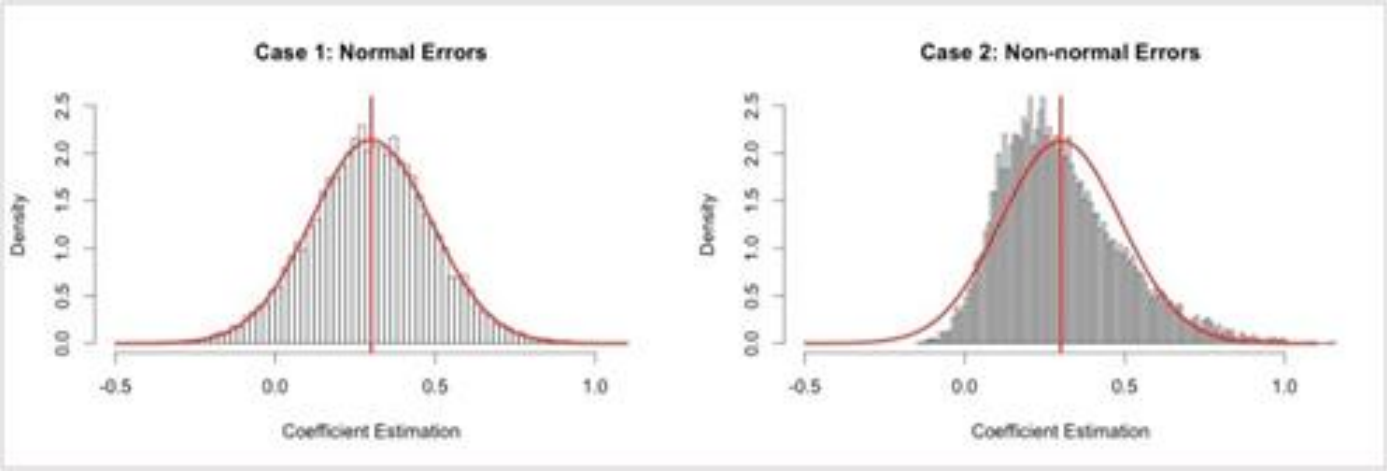
최소제곱추정량이 과소추정된 등분산으로 t-value, F-value 산출

충분히 유의할 수 있는 귀무가설을 기각하는 제 1종 오류 발생

회귀계수 검정에 대한 신뢰성 하락

- **오차의 정규성**

정규분포 자체가 오차에 대한 확률분포이므로 정규분포를 따라야 함



if 가정이 깨진다면?

회귀분석에 사용되는 F-test와 t-test 뿐만 아니라 예측의 경우에도,
모두 오차의 정규성을 가정하므로 결과를 신뢰하기 어려움

- **오차의 독립성**

오차항이 서로 독립일 때를 말함

오차들이 일정한 패턴을 지님
(독립성 위반)

↓

자기상관성 존재

시간적 → **시계열 분석**

공간적 → **공간회귀**

뒤에서 배운다구~

- **오차의 독립성**

  오차항이 서로 독립일 때를 말함

if 가정이 깨진다면?

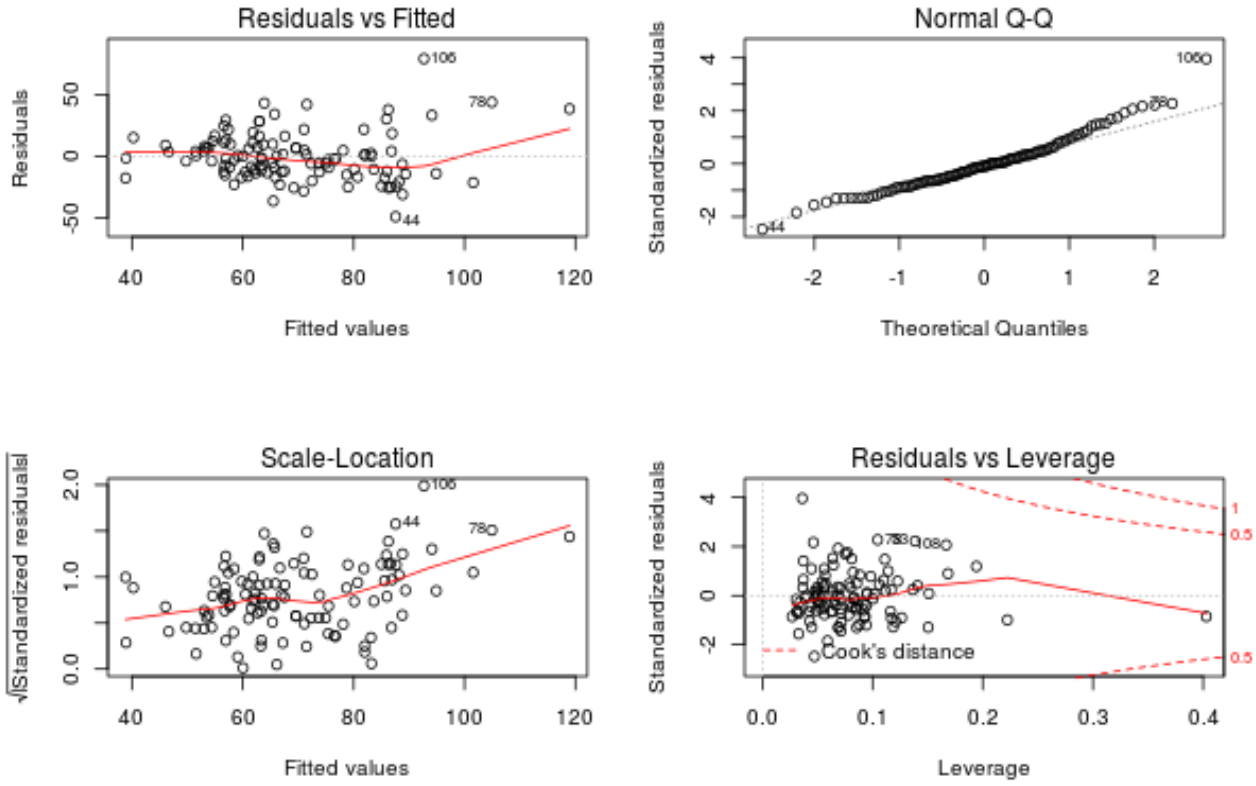
- 최소제곱추정량이 더 이상 BLUE가 아니게 됨
- 분산의 추정량과 회귀계수의 표준오차가 과소추정되어 유의성 검정을 신뢰할 수 없음
- Prediction interval이 넓어짐

2

잔차 플랏

- **잔차플랏**

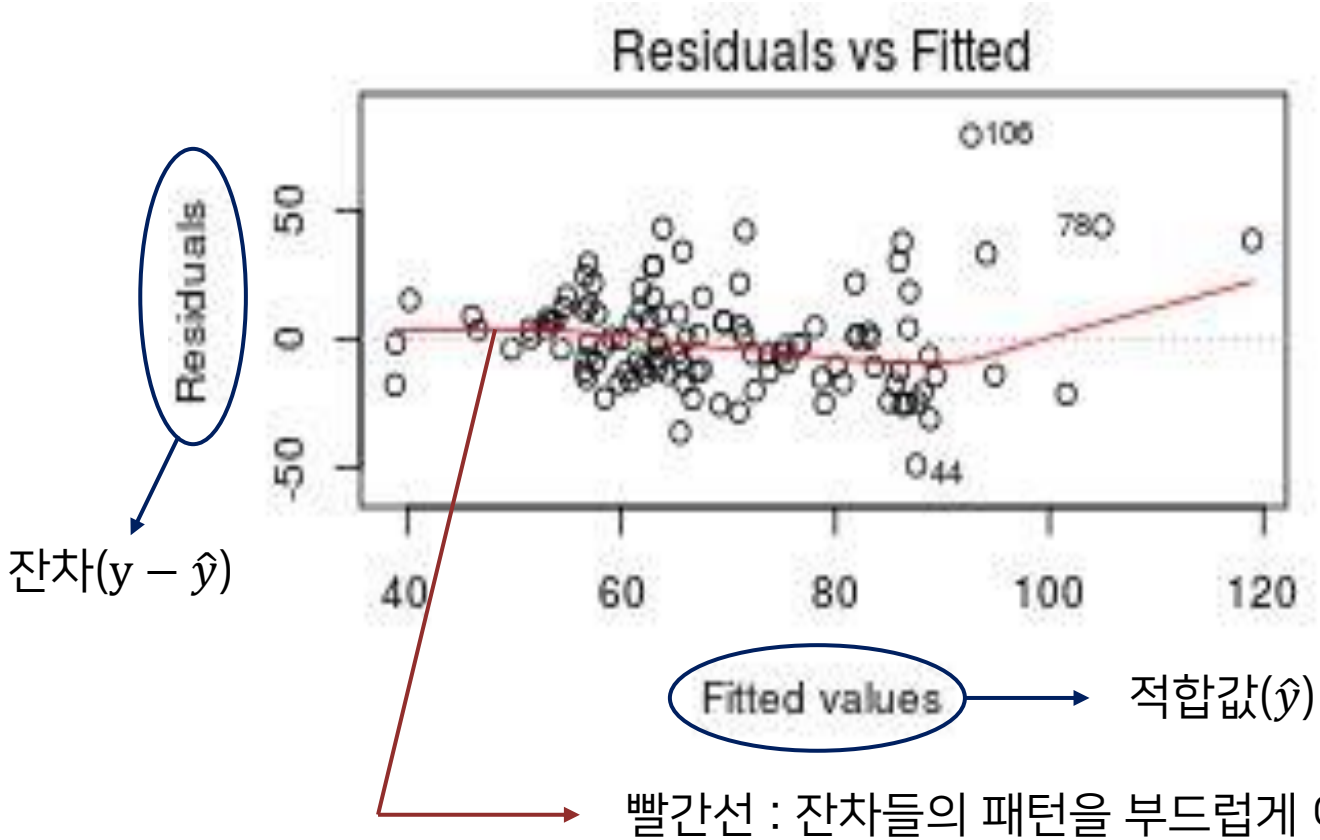
R에서 회귀식을 적합(fitting)하면 제공해주는 네 개의 플랏



회귀모형의 기본가정과 데이터의 문제를 시각적으로 진단 가능!

- Residual vs Fitted

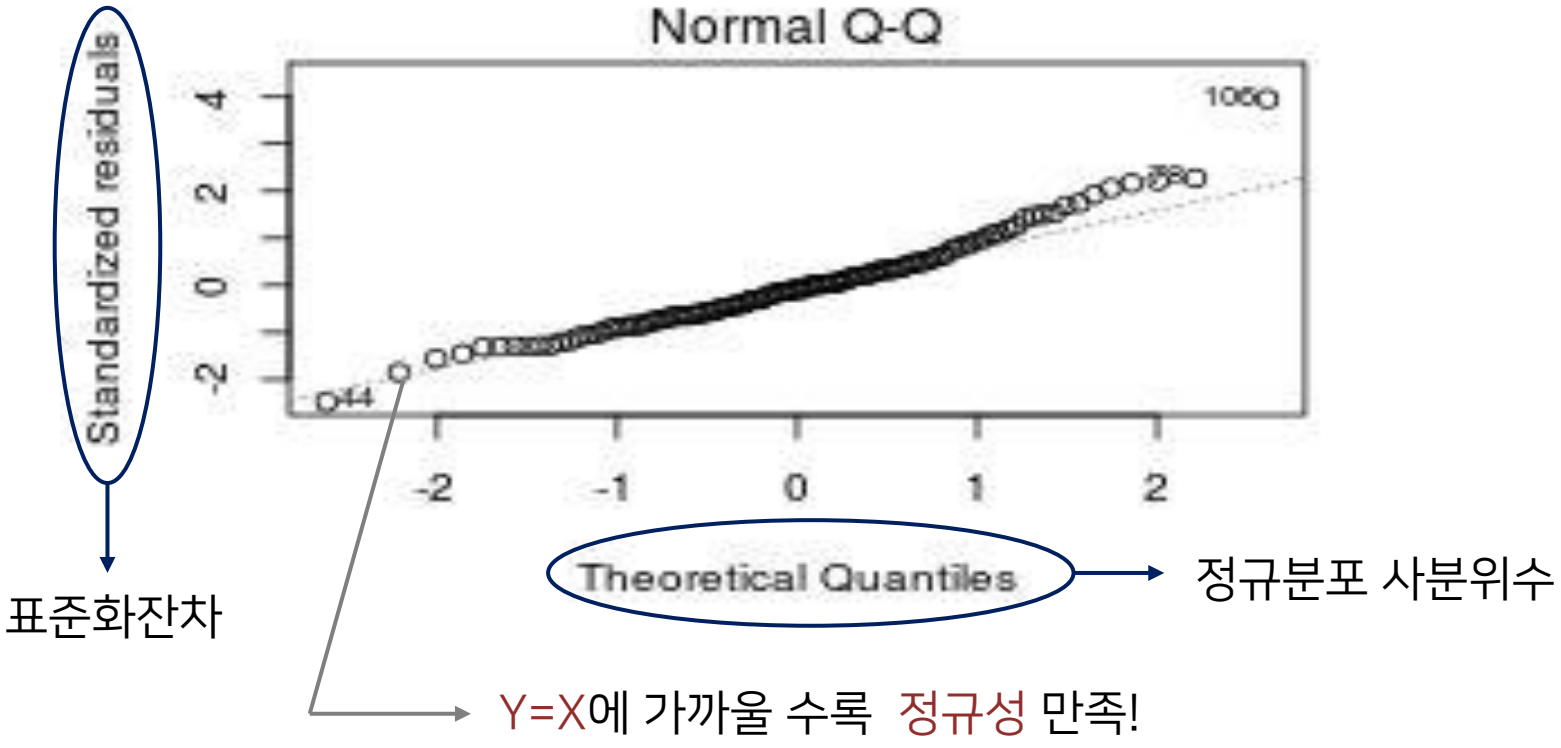
선형성과 등분산성, 독립성 확인이 가능



분포의 추세를 통해 선형성 확인 가능

- Normal QQ plot

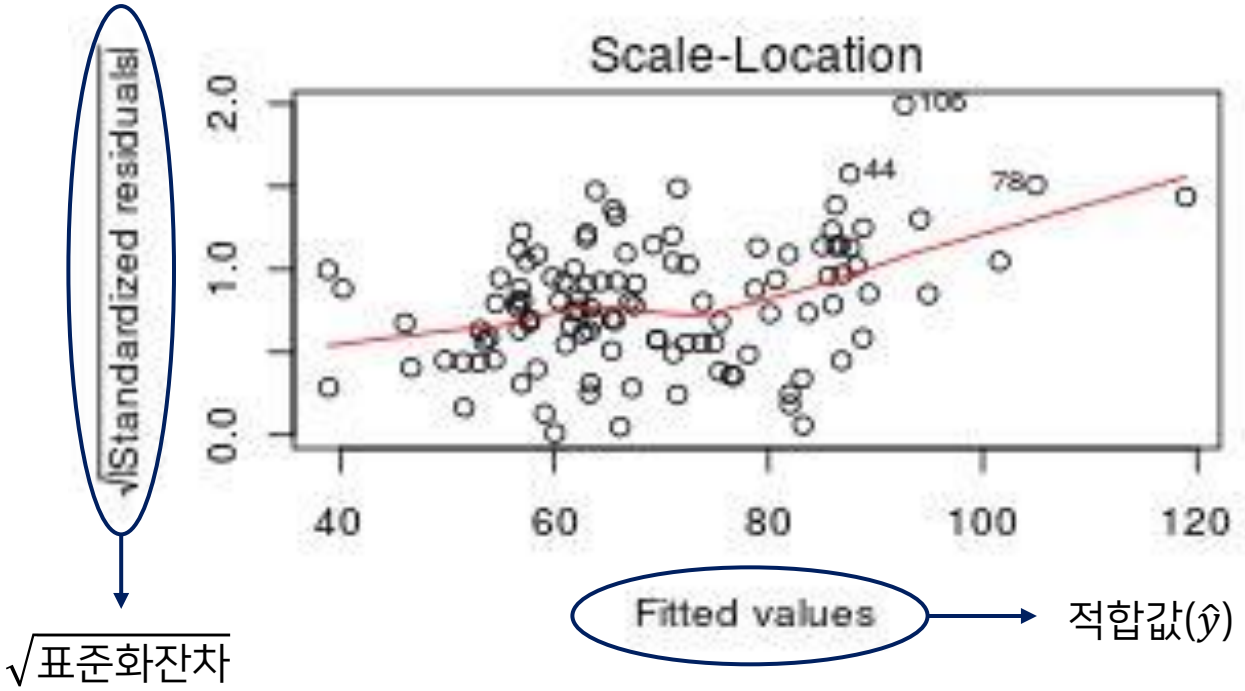
정규성 확인 가능



점들과 점선 간의 관계를 통해 정규성 확인

- Scale-Location

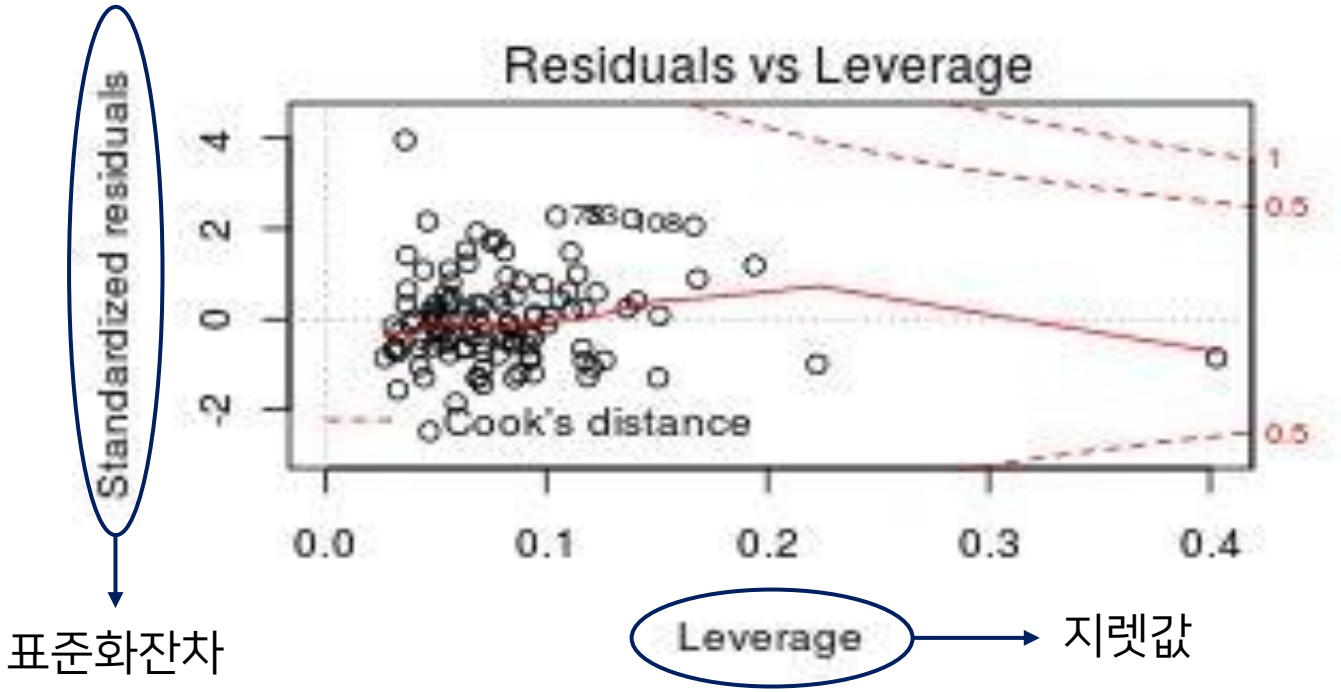
  선형성과 등분산성, 독립성 확인 가능



점의 분포를 통해 등분산성 확인 가능

- Residual vs Leverage

  1주차에 배운 영향점을 파악하기 위한 plot



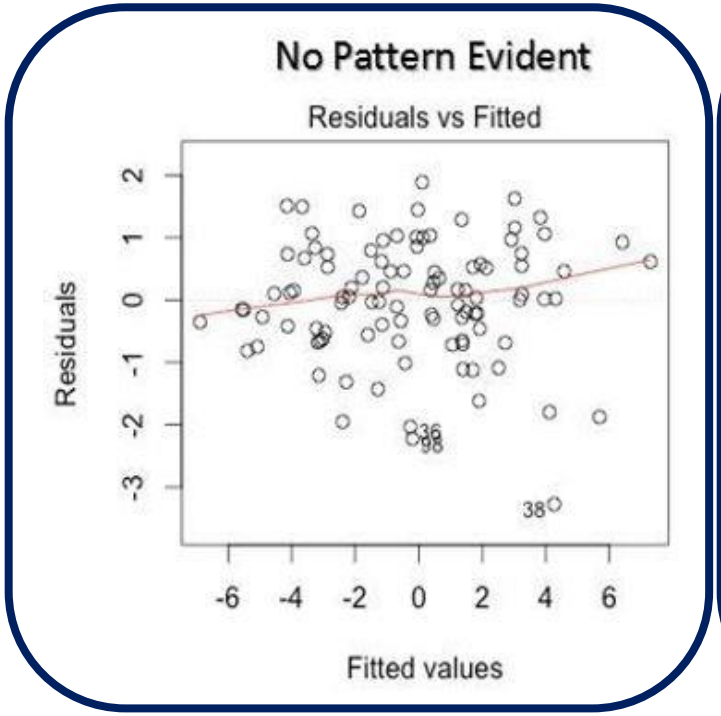
지난 주에 배운 영향점 확인 가능!

3

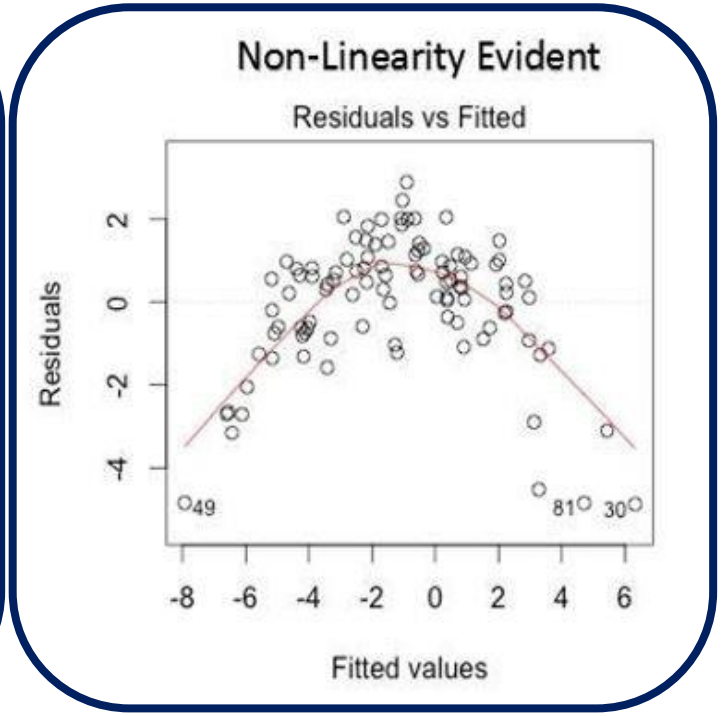
선형성 진단과 처방

- **진단 - 잔차 플랏**

추세선이 X축에 평행한 직선 형태가 아니라면, 선형성 위반!



선형성을 띄는 편



2차 함수 형태이므로
데이터에 대한 추가 설명이 필요함

- 진단 - 잔차 플랏

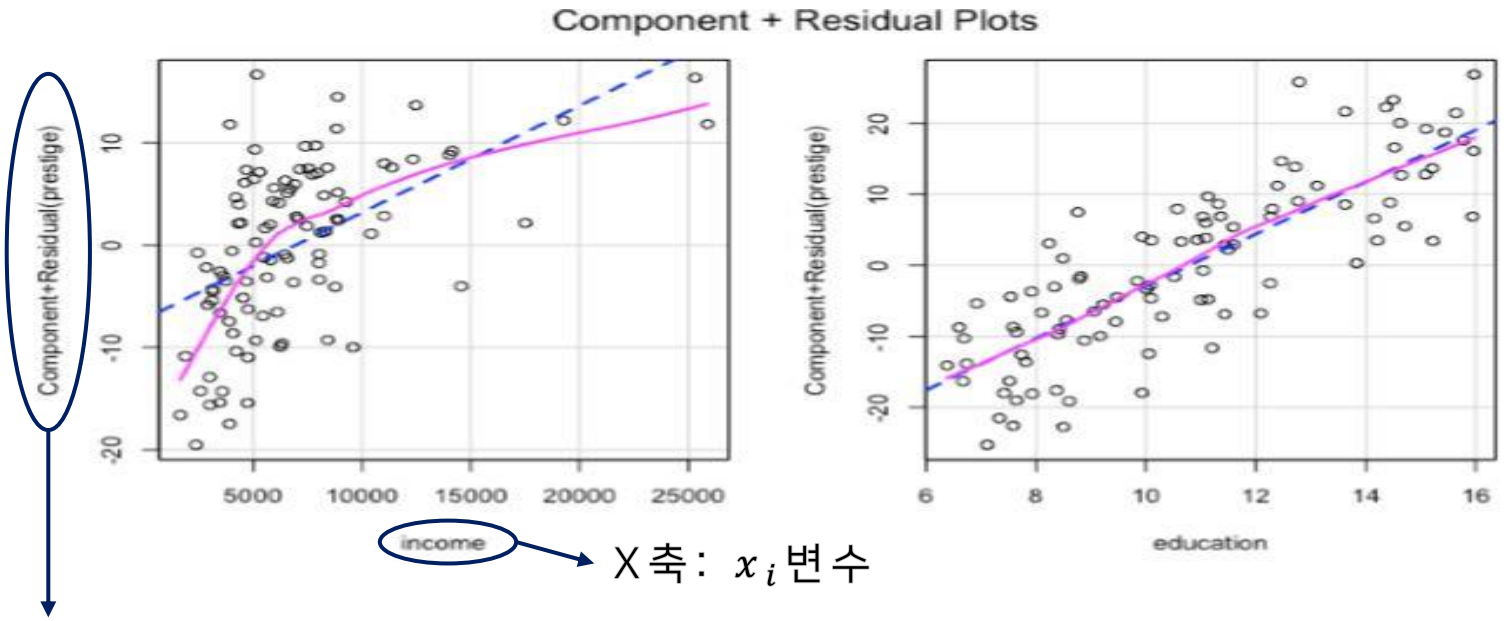
그렇다면, 선형성을 만족하지 못할 때

어떤 변수의 영향으로 인한 것인지 알 수 있을까?

crPlots 을 봐보자!

- 진단 - crPlots

Car 패키지의 CrPlots 함수를 통해 개별 변수의 선형성 파악



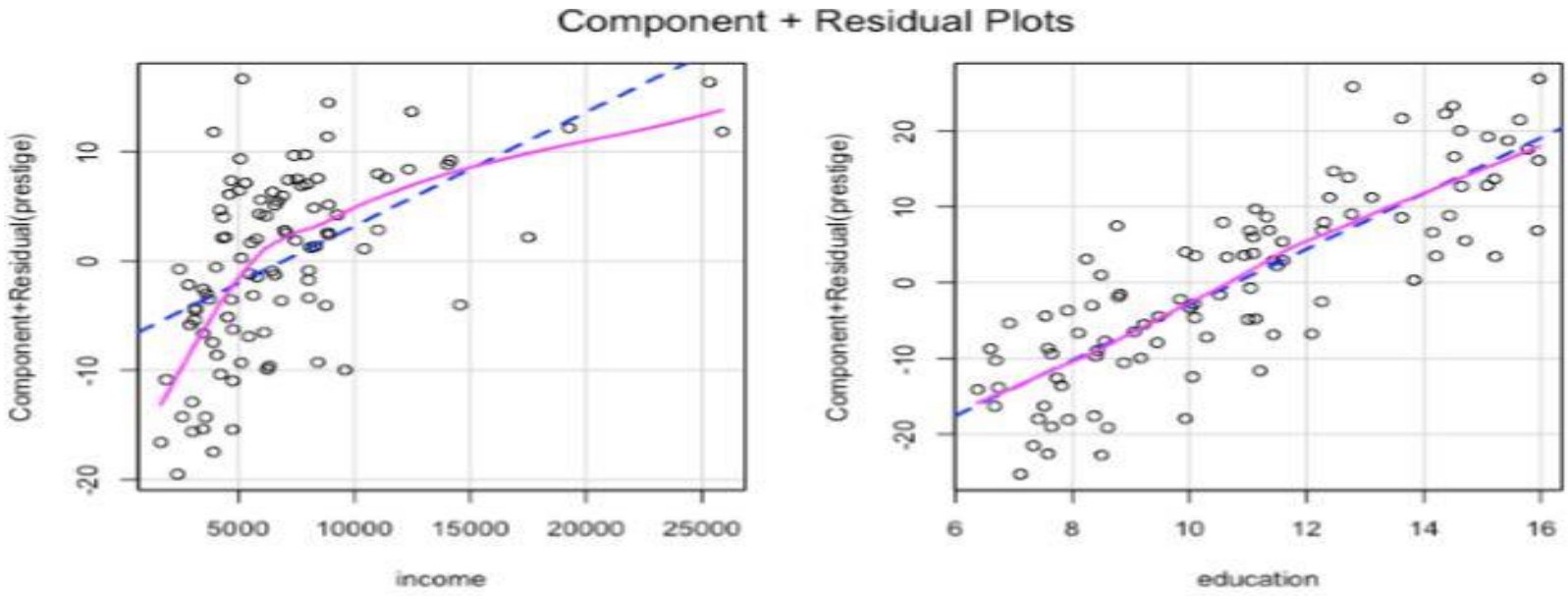
X 축 : x_i 변수

Y 축 : Partial residual ($y - (\beta_i x_i$를 제외한 모든 회귀식 성분))

파란 점선: Partial residual과 x_i의 적합된 직선

보라색 실선: 잔차의 추세선

- 진단 - crPlots



Income 변수가
log형태의 비선형을
잘 잡아내지 못하므로
log 변환 필요!

Education변수가
선형적으로 잘 설명

- crPlots의 한계점

  - X 변수들 사이의 교호작용을 잡아내지 못함

  - 이미 잘못 적합된 변수들이 존재할 경우, 유의미한 관계를 잡아내지 못함

  - 심각한 다중공선성이 존재할 경우 잘못된 정보 제공

3주차 클린업에서 다룰 예정!

3주차도 봐 줄거지?

약속♡

- **처방 - 변수 변환**

비선형 모델의 경우 변수 변환을 통해 해결 가능

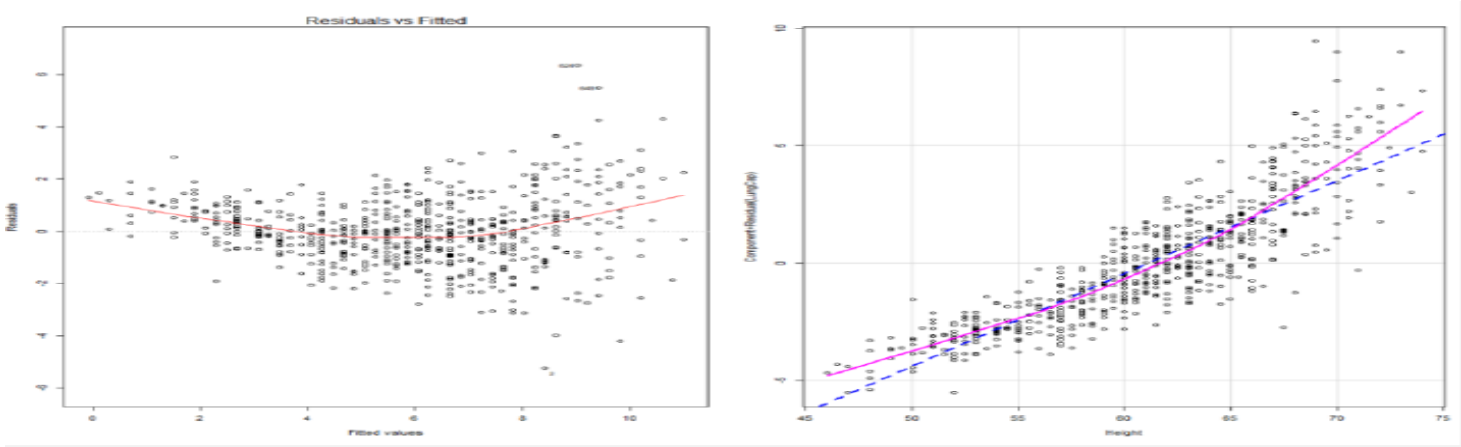
Function	Transformations of x and/or y	Resulting model
$y = \beta_0 x^{\beta_1}$	$y' = log(y), x' = log(x)$	$y' = log(\beta_0) + \beta_1 x'$
$y = \beta_0 e^{\beta_1 x}$	$y' = ln(y)$	$y' = ln(\beta_0) + \beta_1 x$
$y = \beta_0 + \beta_1 log(x)$	$x' = log(x)$	$y = \beta_0 + \beta_1 x'$
$y = \frac{x}{\beta_0 x - \beta_1}$	$y' = \frac{1}{y}, x' = \frac{1}{x}$	$y' = \beta_0 - \beta_1 x'$

X 또는 Y를 변환

선형 관계로 바꿔 줌

- **처방 - Polynomial Regression**

잔차 플랏이나 Partial regression plot을 봤을 때, 2차 이상의 곡선 형태가 나타날 경우에 사용



Residual vs fitted plot
: 2차함수의 모양을 띔

분홍선(polynomial)이
데이터의 2차 모양을 잘 설명

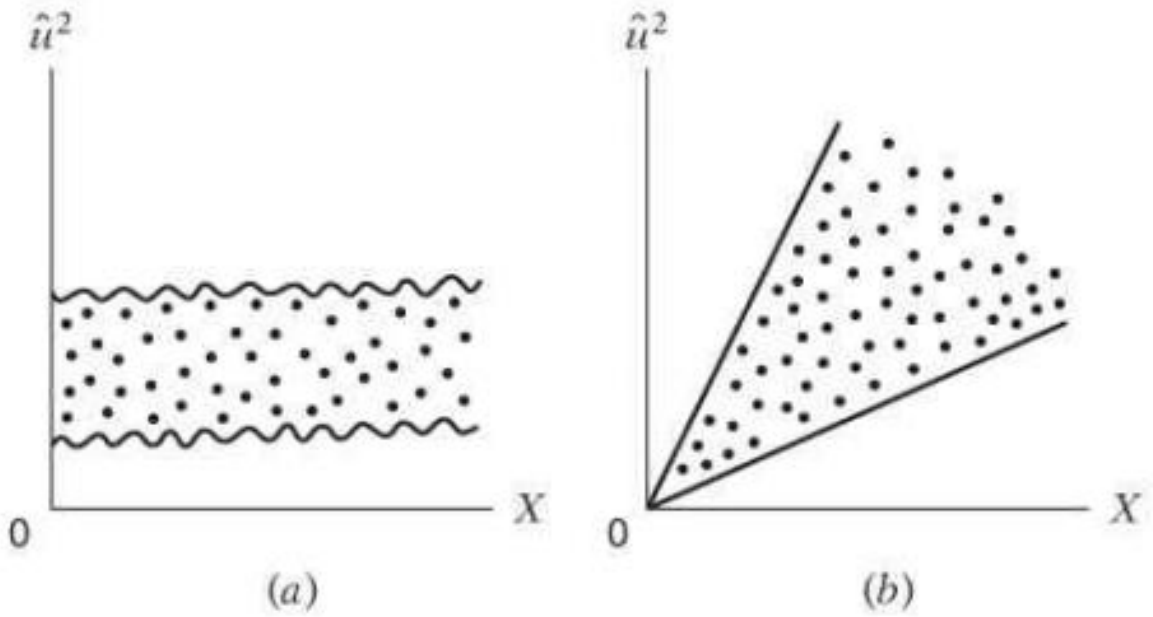
BUT 3차를 넘어서는 모델링은 거의 하지 않음

4

등분산성 진단과 처방

- **진단 – 잔차 플랏**

잔차 플랏 중 'residuals vs fitted' 플랏과 'scale - location' 플랏에서 확인



(a) 등분산성: 표준화 잔차가 0을 기준으로 고르게 분포

(b) 이분산성: 분산이 점점 커지는 모습 등 (a)와 다른 모습

데이터의 퍼짐이 증가·감소·일정하지 않음

- 진단 - test

가설

H_0: 주어진 데이터는 등분산성을 지님

H_1: 주어진 데이터는 등분산성을 지니지 않음

- BP(Breusch-Pagan) test : 가능도비검정 기반으로 추정

- 분산이 설명(X) 변수에 대한 선형결합으로 되어있음 가정

$$\sigma^2 = \alpha_0 + \alpha_1 X_{i1} + \cdots + \alpha_p X_{ip} + u_i$$

- 위의 회귀식의 결정계수(R^2) 값을 구해 검정통계량 계산

$$X^2_{stat} = nR^2 \sim X^2_{p-1}$$

- 진단 - test

가설

H_0: 주어진 데이터는 등분산성을 지님

H_1: 주어진 데이터는 등분산성을 지니지 않음

- BP(Breusch-Pagan) test : 가능도비검정 기반으로 추정

- 단점 : 비선형적인 결합으로 이루어진 이분산성은 파악 불가

샘플이 대표본일 때에만 사용 가능

if $X^2_{stat} > X^2_{p-1,\alpha}$

➡ 귀무가설 기각!

➡ 이분산성 존재

- **처방 - 변수 변환**

  Y를 변환함으로써 등분산 혹은 정규성을 해결해주는 방법

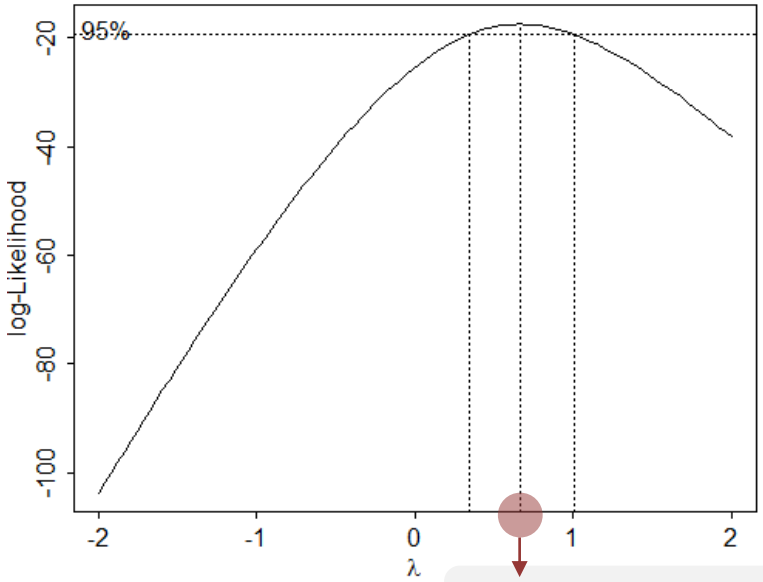
  이때, 통계적인 검정에 따라 구한다는 점에서 효율적

Box-Cox Transformation

Yeo-Johnson Transformation

- Box-Cox Transformation

$$y^{(\lambda)} = \begin{cases} \dfrac{y^{\lambda} - 1}{\lambda}, & if\ \lambda \neq 0 \\ \log y, & if\ \lambda = 0 \end{cases}$$



가능도를 최대로 하는 lambda

- R 코드

```r
#load car library
library(car)

#take a value of lambda
trans = powerTransform(data$variable)
summary(trans)
```

- Y가 0이상일 때에만 사용할 수 있음

- Yeo-Johnson Transformation

$$\psi(\lambda, y) = \begin{cases} \frac{((y+1)^{\lambda}-1)}{\lambda}, & if\ \lambda \neq 0, y \geq 0 \\ \log(y+1), & if\ \lambda = 0, y \geq 0 \\ \frac{-((-y+1)^{2-\lambda}-1)}{2-\lambda}, & if\ \lambda \neq 2, y \leq 0 \\ -\log(-y+1), & if\ \lambda = 2, y \leq 0 \end{cases}$$

- R 코드

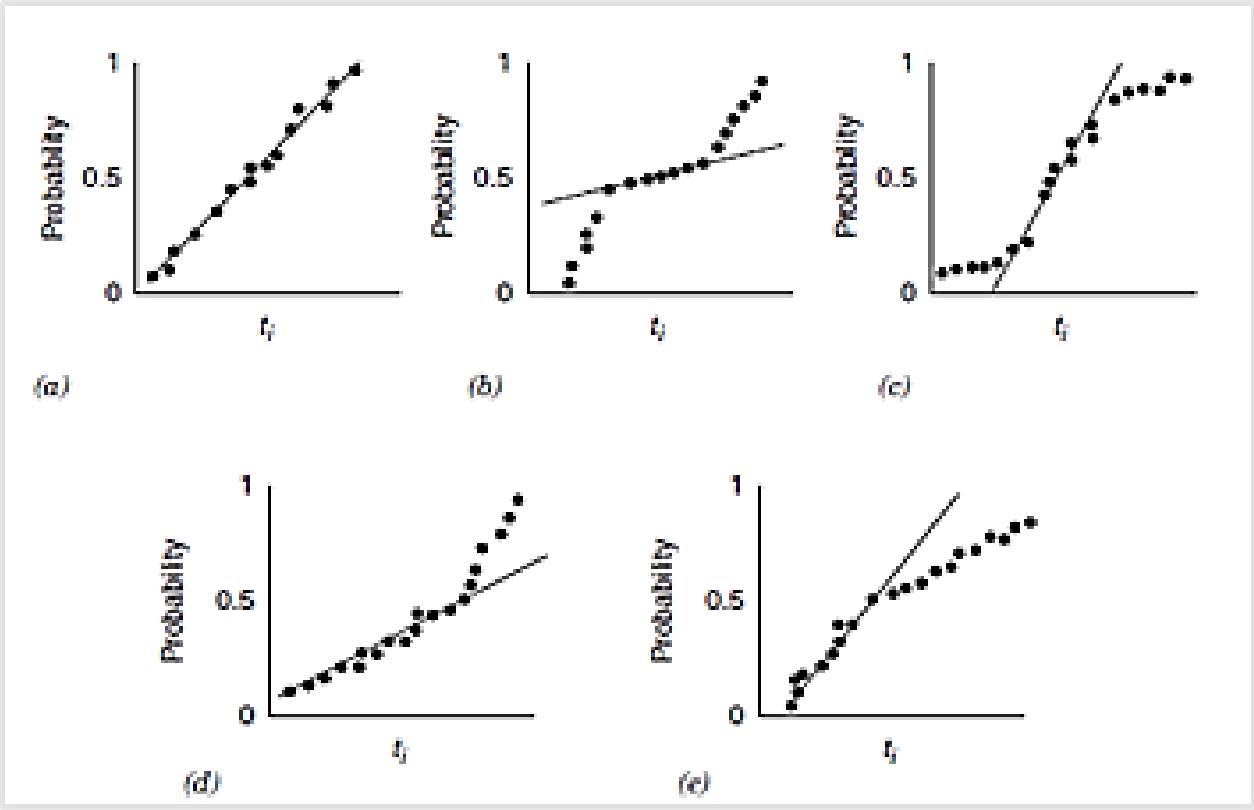
```
#take a value of lambda
trans=powerTransform(data$variable, family="yjpower")
summary(trans)
```

5

정규성 진단과 처방

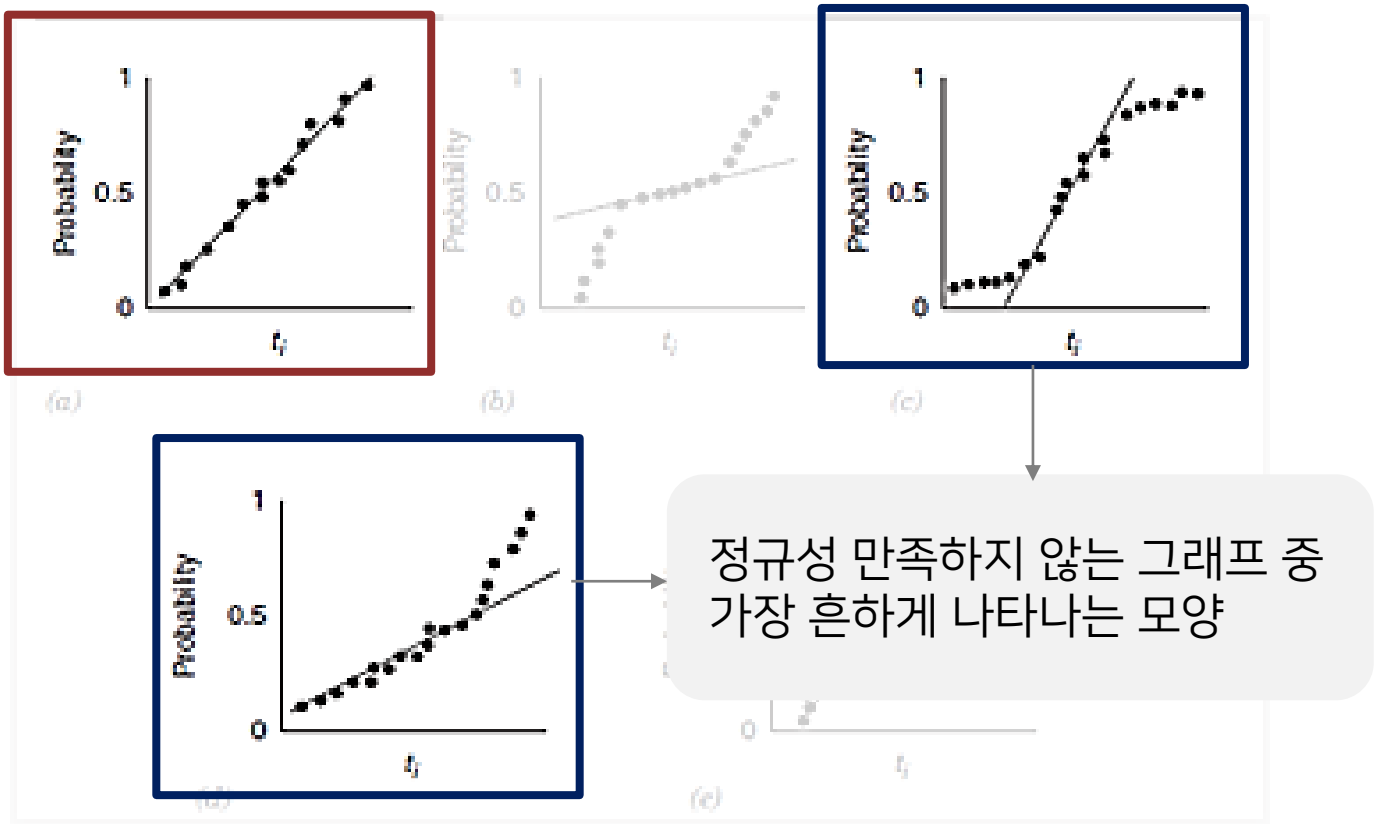
- 진단 - Normal QQ Plot

점들이 $y = x$에 가까우면 정규성 만족



- 진단 - Normal QQ Plot

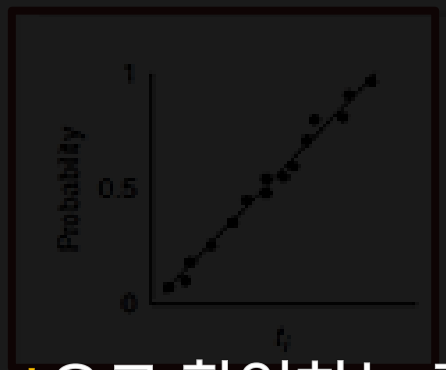
점들이 $y = x$에 가까우면 정규성 만족



Plot 만 보고 결정해도 될까?

- Normal QQ Plot

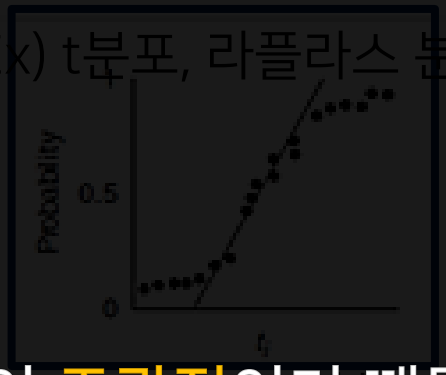
점들이 $y = x$에 가까우면 정규성 만족



NO!!!

Heavy-tail 형태

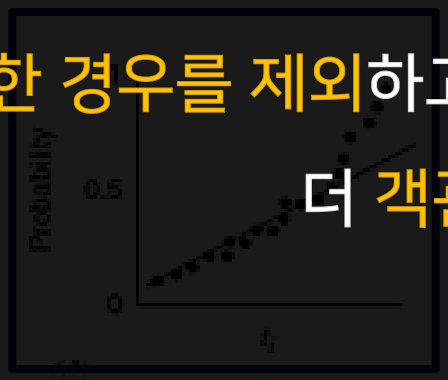
Ex) t분포, 라플라스 분포



Plot으로 확인하는 경우에는 판단이 주관적이기 때문에

너무 명확한 경우를 제외하고는 test를 통해 확인하는 것이

더 객관적이다!



- 진단 - test

가설

H_0: 주어진 데이터는 정규분포를 따름

H_1: 주어진 데이터는 정규분포를 따르지 않음

- Shapiro Wilk Test

  - QQ plot의 아이디어와 동일
  - 정규분포 분위수값(x)과 표준화잔차(y) 사이의 선형관계를 확인하는 검정
  - 관측치가 5000개 이하일 때만 사용 가능

- 진단 - test

가설

H_0: 주어진 데이터는 정규분포를 따름

H_1: 주어진 데이터는 정규분포를 따르지 않음

- Shapiro Wilk Test

R 기본함수로 내장되어 있으며, residual 값을 넣음

```
shapiro.test(salary.reg$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  salary.reg$residuals
## W = 0.96887, p-value = 1.41e-08
```

0.05기준으로 p-value가 작으니 귀무가설 기각
∴ 정규성을 만족하지 않음

- 진단 - test

가설

H_0: 주어진 데이터는 정규분포를 따름

H_1: 주어진 데이터는 정규분포를 따르지 않음

- Jarque-Bera Test

정규분포에서 멀어질수록 통계량 값이 커져 유의수준을 넘어서면 귀무가설 기각

정규분포의 왜도가 0, 첨도가 3이라는 점 기반

$$JB = n\left(\frac{\left(\sqrt{(skew)}\right)^2}{6} + \frac{(kurt-3)^2}{24}\right)$$

- 진단 - test

가설

H_0: 주어진 데이터는 정규분포를 따름

H_1: 주어진 데이터는 정규분포를 따르지 않음

- Jarque-Bera Test

< R 코드 >

```
# Jarque-Bera Test
library(tseries)
jarque.bera.test(fit$residuals)
```

R로 이-지하게 구현 가능!

- **처방 - 변수 변환**

  등분산성과 처방이 같음

  - Box-Cox Transformation
  - Yeo-Johnson Transformation

  * 정규성을 먼저 수정하는 경우가 많음

6

독립성 진단과 처방

- 진단 - test

가설

H_0: 잔차들이 서로 독립 (자기상관성이 없음)

H_1: 잔차들이 서로 독립이 아님 (자기상관성이 있음)

- Durbin Watson Test

바로 앞 뒤 관측치의 자기상관성을 확인하는 테스트

$$d = \frac{\sum_{t=2}^{T}(e_t - e_{t-1})^2}{\sum_{t=1}^{T} e_t^2} \approx 2(1 - \widehat{\rho_1}), \qquad \widehat{\rho_1} = \frac{\sum_{t=2}^{n} e_i e_{i-1}}{\sum_{t=1}^{n} e_t^2}$$

$\widehat{\rho_1}$: 표본 잔차 상관, e_i와e_{i-1}의 상관계수의 꼴, $-1 \le \widehat{\rho_1} \le 1$

$$0 \le d \le 4$$

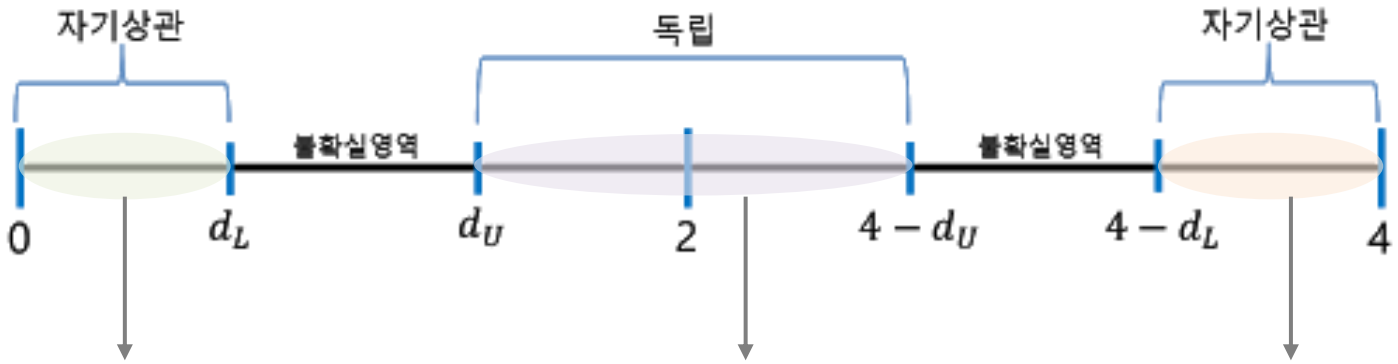
- 진단 - test

가설

H_0: 잔차들이 서로 독립 (자기상관성이 없음)

H_1: 잔차들이 서로 독립이 아님 (자기상관성이 있음)

- Durbin Watson Test



양의 상관관계 (자기상관성 존재) 자기상관성이 없음 음의 상관관계 (자기상관성 존재)

- 진단 - test

가설

H_0: 잔차들이 서로 독립 (자기상관성이 없음)

H_1: 잔차들이 서로 독립이 아님 (자기상관성이 있음)

- Durbin Watson Test

< 단점 >

1) 첫번째 자기 상관성만 알 수 있음 (AR(1)구조만 파악할 수 있음)

2) 자기상관이 오래 지속·계절성이 있는 경우, 확인하는데 한계 존재

- 진단 - test

가설

H_0: 잔차들이 서로 독립 (자기상관성이 없음)

H_1: 잔차들이 서로 독립이 아님 (자기상관성이 있음)

- Durbin Watson Test

< R 코드 >

```
#train a linear model
fit <- lm(y-x1+x2+x3)

#perform Durbin-Watson test1
library(lmtest)
dwtest(fit)
```

- **처방 방법**

1) 가변수 만들기

뚜렷한 계절성이 있다고 판단되면, 이를 위해 가변수를 만듦

2) 시계열 분석

[목차]

시계열팀 교안 보러 갈래요?

7

공간회귀분석

- **공간데이터**

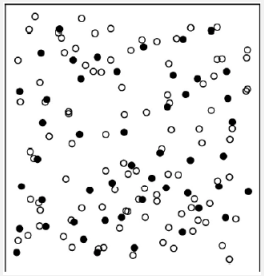
  공간 상의 위치 또는 좌표와 관련된 속성의 집합

공간회귀

- 특정한 현상이 공간상에 분산 또는 집중되었는지 파악

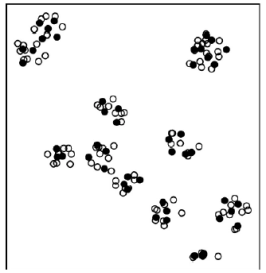
- 공간패턴을 형성하는데 영향을 미친 요인 파악

CSR Pattern



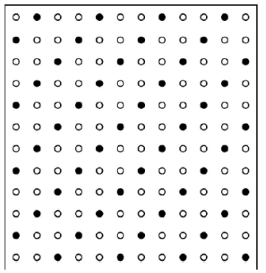
랜덤하게 분포

Cluster Pattern



클러스터링 형성

Decluster Pattern



규칙적으로 분포

- 공간데이터의 특성: 공간자기상관

Tobler 지리학 제 1 법칙

Everything is related to everything else,
but near things are more related than distant things.

공간상에 분포하는 객체 간의 상호작용을 공간자기상관이라 함

강남구은 노원구보다
지리적으로 가까운 서초구와
아파트 가격이 비슷함

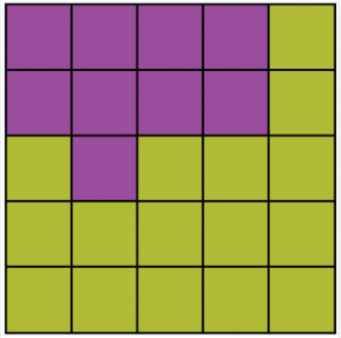
- **공간데이터의 특성: 공간자기상관**

  공간상에 분포하는 객체 간의 상호작용

정적 공간자기상관

Positive pattern of similarity

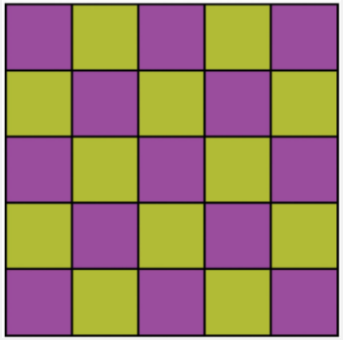
: 공간 객체 값이 근처의 객체 값과 유사한 값을 가지는 것



부적 공간자기상관

Negative pattern of similarity

: 공간 객체 값이 근처의 객체 값과 상반된 값을 가지는 것



- **공간데이터의 특성: 공간자기상관**

공간상에 분포하는 객체 간의 상호작용

전역적 공간자기상관
Global Spatial Autocorrelation

전체 구역이 가지는
하나의 공간자기상관의 정도

예시 한국 전체에서 나타나는
고혈압 유병률의 공간적인 패턴

국지적 공간자기상관
Global Spatial Autocorrelation

개별 지점이 가지는
공간자기상관의 정도

예시 수도권에서 나타나는
고혈압 유병률의 공간적인 패턴

공간 자기상관은…

특정 지역의 사건 강도가 인접 지역의 사건에 영향을 주는지를 파악하자!

- **공간데이터의 특성: 공간적 이질성**

  넓은 지역에서 나타나는 불규칙한 분포를 의미하며,

  한 지역 내에 서로 다른 성격의 하위 집단이 존재하는 것을 말함

Example

지하철 개통이 지가에 미치는 영향력의 크기가 모든 지역에서 같은가?

→ 영향을 크게 받는 지역, 영향을 크게 받지 않는 지역 등 여러 유형의 집단 존재

공간적 이질성은…

특정 사건이 전 지역에서 동일한 강도로 나타나는지 파악하자!

- **공간자기상관 진단**

먼저 지역 내 지점들이 인접해 있는지부터 체크해야 한다!

공간가중행렬(Spatial Weights Matrix)

- 지역 내 다수의 지점들이 서로 공간적으로 인접하고 있는지의 여부를 파악할 수 있도록 행렬로 나타낸 것

- 지역 간의 잠재적 상호작용의 강도를 나타냄

$$w_{ij} = \begin{cases} 1 & \text{if } i,j \text{ is } neighbor \\ 0 & otherwise \end{cases}$$

이웃? 기준이 뭔데?

- 공간가중행렬의 이웃 결정 기준

<table>
<tr><td rowspan="3">Binary Contiguity Weights</td><td>Bishop Contiguity</td></tr>
<tr><td>Rook Contiguity</td></tr>
<tr><td>Queen Contiguity</td></tr>
<tr><td colspan="2">Distance-based Weights</td></tr>
<tr><td colspan="2">K-Nearest Neighbors Weights</td></tr>
</table>

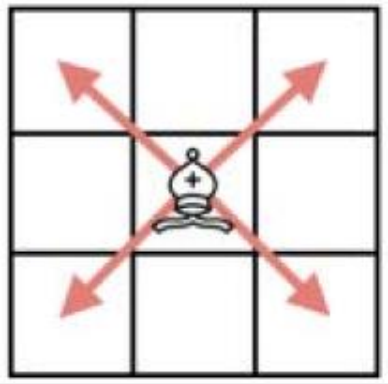
다음 슬라이드에서
하나씩 살펴보자…!

- 공간가중행렬의 이웃 결정 기준

Binary Contiguity Weights

: 근접하고 있는 경우를 이웃으로 보는 방법

Bishop Contiguity



각 모서리에 있는 영역을

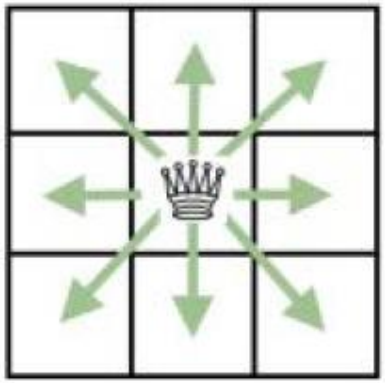
이웃으로 간주

- 공간가중행렬의 이웃 결정 기준

Binary Contiguity Weights

: 근접하고 있는 경우를 이웃으로 보는 방법

Queen Contiguity



면이나 모서리가

붙어있는 영역을

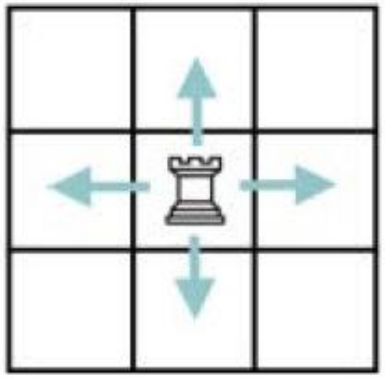
모두 이웃으로 간주

- 공간가중행렬의 이웃 결정 기준

Binary Contiguity Weights

: 근접하고 있는 경우를 이웃으로 보는 방법

Rook Contiguity



각 면이 붙어있는 영역을

이웃으로 간주

가장 보편적!

Rook Contiguity를 사용하여 직접 공간가중행렬을 만들어보자!

A	B	C
D	E	F
G	H	I

E를 기준으로 B, D, F, H가 면이 붙어 있으므로 이웃으로 간주

$$W = \begin{array}{c} \begin{matrix} A & B & C & D & E & F & G & H & I \end{matrix} \\ \begin{pmatrix} 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \end{pmatrix} \end{array} \begin{matrix} A \\ B \\ C \\ D \\ E \\ F \\ G \\ H \\ I \end{matrix}$$

- **공간가중행렬의 이웃 결정 기준**

Distance-based Weights

: 특정 거리보다 가까우면 이웃이라고 정의하는 방법

Distance-based Weights

$$w_{ij} = \begin{cases} 1 & if \;\; d_{ij} < d \\ 0 & otherwise \end{cases}$$

d를 너무 작게 잡으면 이웃이 없는 고립된 점이 생길 수 있으므로,

d를 각 관측치별 최단거리보다는 크게 정해야 함

- 공간가중행렬의 이웃 결정 기준

K-Nearest Neighbors Weights

: 머신러닝의 KNN 방법과 비슷한 방식으로, 가장 근접한 K개의 점을 이웃으로 정의하는 방법

이렇게 만들어진 공간가중행렬은
그대로 쓰이지 않고, 정규화하여 사용됨

끝난줄 알았지?

- 공간가중행렬의 정규화 방법

Row Standardized Weights

: 행 단위로 정규화하는 방법으로, 가중치를 각 행의 합으로 나눠줌

$$\begin{pmatrix} 1 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix} \Rightarrow \begin{pmatrix} 1/3 & 1/3 & 1/3 \\ 1/2 & 0 & 1/2 \\ 1/2 & 1/2 & 0 \end{pmatrix}$$

- 공간가중행렬의 정규화 방법

Stochastic Weights

: 전체 행렬을 정규화하는 방법으로, 가중치를 행렬 전체의 합으로 나눠줌

$$\begin{pmatrix} 1 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix} \Rightarrow \begin{pmatrix} 1/7 & 1/7 & 1/7 \\ 1/7 & 0 & 1/7 \\ 1/7 & 1/7 & 0 \end{pmatrix}$$

이렇게 정규화를 마친 후, 공간자기상관성에 대한 검정 시행!

- **공간자기상관 진단**

  - Moran's I 지수

  : 지역 간의 인접성을 나타내는 공간가중행렬과 인접하는 지역들 간의 속성 데이터의 유사성을 측정하는 방법

$$I = \frac{N \sum_i^N \sum_j^N w_{ij}(Y_i - \bar{Y})(Y_j - \bar{Y})}{(\sum_i^N \sum_j^N w_{ij}) \sum_i^N (Y_i - \bar{Y})^2}$$

$$Z_I = \frac{I - E(I)}{\sqrt{Var(I)}} \quad where \;\; E(I) = -\frac{1}{N-1}$$

N: 지역단위수, Y_i: i 지역의 속성, Y_j : j 지역의 속성, $\bar{Y}$: 평균값, w_{ij} : 가중치

수식은
가볍게 패스하자

- **공간자기상관 진단**

  - Moran's I 지수

  : 지역 간의 인접성을 나타내는 공간가중행렬과 인접하는 지역들 간의 속성 데이터의 유사성을 측정하는 방법

Moran's I Index

- Z_I 값이 검정통계량이며, Z검정을 통해 전역적 공간자기상관성의 통계적 유의성 판단
- 전체 공간에서 패턴이 있는지 없는지만 알 수 있을 뿐, 핫스팟이나 콜드스팟의 위치는 알 수 없음

* 핫스팟: 사건이 집단화되어 나타나는 특정 지역
* 콜드스팟: 해당지역이 주변지역에 비해 차이성이 큰 지역

- 공간자기상관 진단

  - Moran's I 지수

  : 지역 간의 인접성을 나타내는 공간가중행렬과 인접하는 지역들 간의 속성 데이터의 유사성을 측정하는 방법

전역적 공간 자기상관성이 있다면

세부적으로는 어떤 지역에서

그러한 패턴이 나타날까?

Moran's I Index

- Z_I 값이 검정통계량이며, Z검정을 통해 전역적 공간자기상관성의 통계적 유의성을 검정

- 전체 공간에서 패턴이 있는지 없는지만 알 수 있을 뿐, 핫스팟이나 콜드스팟의 위치는 알 수 없음

* 핫스팟: 사건이 집단화되어 나타나는 특정 지역
* 콜드스팟: 해당지역이 주변지역에 비해 차이성이 큰 지역

- **공간자기상관 진단**

  - LISA 지표

  : 특정 지역들이 전체 지역의 공간 자기상관성에 얼마나 영향을 미치는지 파악하는 국지적 측정 방법

$$I_i = \frac{N^2}{\sum_i^N \sum_j^N w_{ij}} \frac{(Y_i - \bar{Y}) \sum_j^N w_{ij}(Y_j - \bar{Y})}{\sum_i^N (Y_i - \bar{Y})^2}$$

수식은 넘기라고~

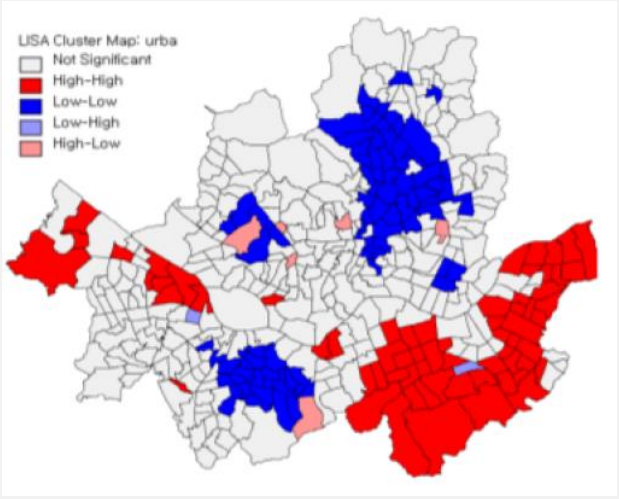
공간 자기상관이 세부적으로 어느 지역에서 나타나는 것인지 알 수 있다!

- **공간자기상관 진단**

  - LISA 지표

  : 특정 지역들이 전체 지역의 공간 자기상관성에 얼마나 영향을 미치는지 파악하는 국지적 측정 방법

LISA clustering Map



HH(high-high), LL(low-low): 공간적 군집지역
HL(high-low), LH(low-high): 공간적 이례지역

공간적 클러스터 패턴이
어떻게 나타나는지 분석 가능

HH,LL: 양의 전역 공간자기상관성에 기여

HL,LH: 음의 전역 공간자기상관성에 기여

회색: 전역 공간자기상관성에 기여 X

- **공간자기상관 진단**

  - 라그랑지 승수검정(LM: Lagrange Multiplier)

  : OLS 회귀모델의 종속변수 또는 오차에서 공간자기상관이 실재하지 않는다는 귀무가설에 대해 검정하는 것

공간회귀모델에서 공간자기상관이

종속변수에서 나타나는지, 오차에서 나타나는지에 따라

사용하는 공간회귀모델이 달라짐

공간회귀모델 선택 알고리즘

모란 I 지수, LISA로 공간 자기상관성이 있는지 확인

라그랑지승수검정으로 모델 선택

유의 X

OLS 회귀모델

LM-Lag 유의

공간시차모델

LM-Error 유의

공간오차모델

둘 다 유의

Robust LM

한 번 더 검정해서

더 유의한 모델 선택

- **공간자기상관 처방**

공간 자기상관성	공간시차모델(SLM)
	공간오차모델(SEM)
공간적 이질성	지리가중회귀모형(GWR)

공간 자기상관성 → 인접지역의 영향력을 변수에 포함시켜 통제

공간적 이질성 → 각 지역마다 다른 추정계수로 영향력을 추정

- **공간시차모델(SLM, Spatial Lag Model)**

한 지역의 관측치가 인접지역의 관측치와 상관성이 있는 경우,
공간적 의존성을 하나의 설명변수로 둔 모델

공간시차변수

$$Y = \rho WY + X\beta + \varepsilon$$
$$= (1 - \rho W)^{-1}(X\beta + \varepsilon),$$

$$\varepsilon \sim MVN(0, \sigma^2 I_n)$$

Spatial Lag Model

주택가격 = 주택면적 + 건축년도 + 가구주의 소득 + 오차

공간시차변수 추가

주택가격 = W*주택가격 + 주택면적 + 건축년도 + 가구주의 소득 + 오차

- **공간오차모델(SEM, Spatial Error Model)**

오차에 공간자기상관성이 있다면, 이는 주로 설명변수를 고려하지 못하였기 때문

→ 오차를 공간오차변수로 만들어 줌

$$Y = X\beta + \mu$$
$$= X\beta + (1 - \lambda W)^{-1},$$

공간오차변수

$$where\ \ \mu = \lambda W\mu + \varepsilon\ \ and$$
$$\varepsilon \sim MVN(0, \sigma^2 I_n)$$

Spatial Error Model

주택가격 = 주택면적 + 건축년도 + 가구주의 소득 + 오차

↓ *공간오차변수로 변형*

주택가격 = 주택면적 + 건축년도 + 가구주의 소득 + 오차(W*오차 + ε)

- **지리가중회귀모델(GWR, Geographically Weighted Regression)**

  변수들 간의 관계를 추정하는 회귀계수가 지역마다 서로 다른 것을 전제로 지역별로 회귀모델을 추정하는 방법

$$W_i^{1/2}Y = W_i^{1/2}X\beta_i + W_i^{1/2}X\varepsilon_i$$

$$\beta(u_i, v_i) = [X'W(u_i, v_i)X]^{-1}X'W(u_i, v_i)XY$$

추정된 계수값은 해당 격자에서만 의미 있음

수식은 패스 ^^

- **W** : 공간의 특성을 반영하고 있는 변수가 거리에 따라 얼마나 민감하게 변하는지 보여주는 지표

- 각 위치좌표 (u_i, v_i)별로 하나씩의 W를 가지며, W는 인근 관측치로부터 도출됨

! 주의 ! W 공간가중행렬 아님!!!

$$W_i^{1/2} Y = W_i^{1/2} X\beta_i + W_i^{1/2} X\varepsilon_i$$

$$\beta(u_i, v_i) = [X'W(u_i, v_i)X]^{-1} X'W(u_i, v_i)XY$$

- **지리가중회귀모델(GWR, Geographically Weighted Regression)**

  - W 만드는 방법

Exponential 가중치

$$W_i = \sqrt{\exp(-d_i/\theta)}$$

d_i: i 지역에서부터 다른 지역까지의 거리,

θ: 대역폭

- 대역폭 값과 거리에 대한 가중치 값은 비례

- 대역폭 선택의 영향을 가장 많이 받음

대역폭은 CV를 통해 튜닝할 수 있다!

지리가중회귀모델(GWR) 사용시 주의사항

- 지리가중회귀모델(GWR, Geographically Weighted Regression)

공간적 이질성이 있을 때 GWR을 사용하지만,
공간적 이질성을 판단하는 정확한 방법이 없다!

GWR을 사용했을 때, 지역별로 같은 변수에 대한 회귀계수의 차이가 크다면 공간적 이질성을 고려해볼 수 있다.

GWR 사용 후에는 반드시 F검정을 이용해 대안모형(GWR)이 기준모형을 개선했는지 점검!

3주차 예고

1. 다중공선성
2. 변수선택법
3. 축소 추정

THANK YOU

8

부록

- 진단 - test

가설

H_0: 주어진 데이터는 등분산성을 지님

H_1: 주어진 데이터는 등분산성을 지니지 않음

- NCV(Non-Constant Variance) score test: score test 기반으로 추정

  - BP test의 일반화된 버전

  - 표준화되지 않은 BP test와 유사하지만 완전히 동일하지 않음

- **처방 - WLS(Weighted Least Square)**

데이터마다 다른 가중치를 주어 등분산을 만족하게 하는 일반화된 최소제곱법

분산이 커 신뢰도가 낮은 부분의 관측치에는 작은 가중치

분산이 작아 신뢰도가 높은 부분의 관측치에는 큰 가중치

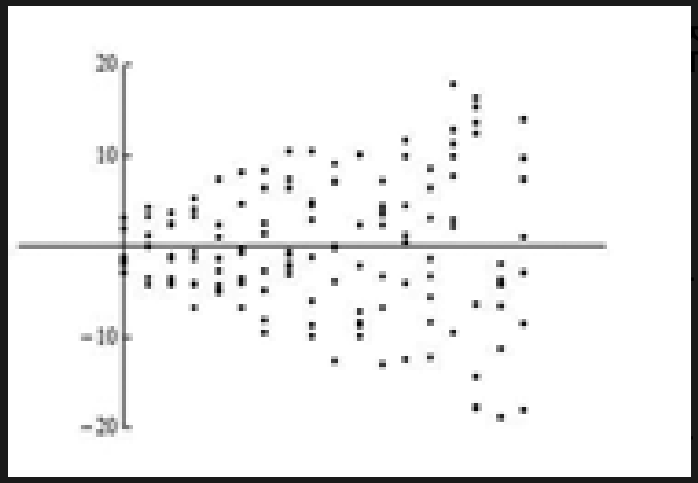
$$\sum \omega_i \left(y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip}\right)^2, \qquad \omega_i = \frac{1}{\sigma_i^2}$$

분산을 우리가 알기 어렵기 때문에
가중치는 경험적으로 선정해야함

WLS를 통해 구한 추정량은 BLUE!

WLS의 가중치를 정하는 방법

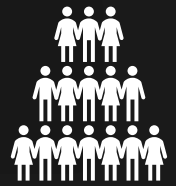
- 잔차 플랏
- 사전 지식 이용



- 반응변수가 평균 값일 때
  각 표본의 크기가 다른 경우

- 표본크기(n_i)가 클 수록
  y_i의 신뢰도가 높아지므로 $\omega_i = n_i$

위의 그래프처럼
분산이 점점 커질 경우

$\omega_i = 3$  $\omega_i = 100$ 

3명(n_i)　　100명(n_i)

가중치는 $\omega_i = \frac{1}{x_i^2}$

- 진단 - test

가설

H_0: 주어진 데이터는 정규분포를 따름

H_1: 주어진 데이터는 정규분포를 따르지 않음

- Anderson-Darling Test

  - Empirical CDF(경험적 누적밀도함수)를 통해 검정
  - 정규분포의 확률밀도함수, 누적밀도함수의 형태와 유사한지를 검정
  - R 코드

```
#Anderson-Darling Test
library(nortest)
ad.test(fit$residuals)
```

- **Gvlma package (Global Validation of Linear Model Assumption)**

  선형성, 정규성, 등분산성을 한 번에 체크해주는 함수

- Global Stat

  선형성 체크, $p<0.05$이면 귀무가설 기각

- Skewness, Kurtosis

  정규성 체크, $p<0.05$이면 귀무가설 기각

- Heteroscedasticity

  등분산성 체크, $p<0.05$이면 귀무가설 기각

- **Gvlma package (Global Validation of Linear Model Assumption)**

  선형성, 정규성, 등분산성을 한 번에 체크해주는 함수

```
Call:
 gvlma(x = salary.reg)

                     Value   p-value                   Decision
Global Stat        113.688 0.000e+00 Assumptions NOT satisfied!
Skewness            37.022 1.168e-09 Assumptions NOT satisfied!
Kurtosis            50.181 1.402e-12 Assumptions NOT satisfied!
Link Function       25.760 3.867e-07 Assumptions NOT satisfied!
Heteroscedasticity   0.725 3.945e-01    Assumptions acceptable.
```

$3.9452 \times e^{-1} \approx 1.451 \geq 0.05$

∴ 귀무가설을 기각하지 못함

∴ 등분산성은 만족하는 것으로 보임

- **공간데이터**

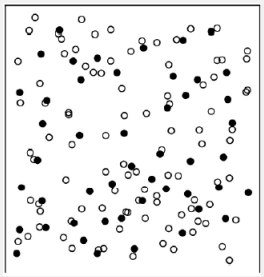
  공간 상의 위치 또는 좌표와 관련된 속성의 집합

OLS 회귀

Spatial Randomness 가정

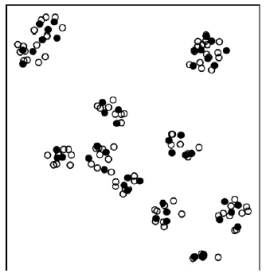
공간상에서 어떠한 사건이 발생할 확률의 분포는 전부 같으며, 아무런 패턴이 없다.

CSR Pattern



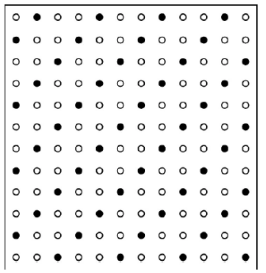
랜덤하게 분포

Cluster Pattern



클러스터링 형성

Decluster Pattern



규칙적으로 분포

- **지리가중회귀모델(GWR, Geographically Weighted Regression)**

  - W 만드는 방법

Tricube 가중치

$$W_i = \begin{cases} (1-(d_i/q_i)^3)^3 & where\ d_i < q_i \\ 0 & otherwise \end{cases}$$

q_i: 지역 i로부터 q개만큼 인접한 지역까지의 거리

거리에 대한 가중치값은 q개만큼 인접한 지역까지의 거리에 의해 결정되도록 구성

Gaussian 가중치

$$W_i = \phi(d_i/\sigma\theta)$$

$\phi(\cdot)$: 표준정규분포함수의 pdf